Introduction to Machine Translation

Dr. Mariana Neves

October 17th, 2016
Overview

- Introduction
- Applications
- Challenges
- History
- Available resources
- MT paradigms
- MT course
Overview

- Introduction
- Applications
- Challenges
- History
- Available resources
- MT paradigms
- MT course
Machine translation (MT)

- Automatic translation from one language to another
- Koehn: „Translating between languages is [...] a task for which even humans require special training.“
Machine translation

Transfer market: Three Dortmund shoot in market value after the top.
As she lay there alone Daiyu's thoughts turned to Baochai.

The coldness penetrated the curtains of her bed.

Almost without noticing it she had began to cry.
Overview

- Introduction
- Applications
- Challenges
- History
- Available resources
- MT paradigms
- MT course
MT applications

- Assimilation
  - Understand the content

- Dissemination
  - Publication in other languages

- Communication
  - Emails, chats

[Koehn 2010]
Fully Automatic High Quality Machine Translation (FAHQMT)

- Limited domains (weather, sport, rail, flight info)
- Controlled vocabulary
Météo: translation of weather reports and warnings

Since 2004, RALI has been investigating how well current SMT approaches deal with a real-world task. We have reconstructed translation systems for dealing with weather bulletins and warnings issued by the Canadian Meteorological Centre of Environment Canada.

<table>
<thead>
<tr>
<th>#</th>
<th>ED</th>
<th>PER</th>
<th>REP</th>
<th>Source</th>
<th>SMT</th>
<th>REF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.000</td>
<td>21</td>
<td>THESE THUNDERSTORMS WILL PRODUCE GUSTY WINDS OF 90 KM / H OR MORE, HAIL STONES OF 2 CM OR MORE, HEAVY RAIN AND FREQUENT LIGHTNING.</td>
<td>CES ORAGES PRODUIRONT DES RAFALES DE 90 KM / H OU PLUS - DE LA GRELE DE 2 CM OU PLUS - DE FORTES PLUIES - ET DE NOMBREUX ECLAIRS.</td>
<td>CES ORAGES PRODUIRONT DES RAFALES DE 90 KM / H OU PLUS - DE LA GRELE DE 2 CM OU PLUS - DE FORTES PLUIES - ET DE NOMBREUX ECLAIRS.</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.000</td>
<td>38</td>
<td>PERSONS IN THESE REGIONS SHOULD TAKE SAFETY PRECAUTIONS AND LISTEN FOR SUBSEQUENT WARNINGS.</td>
<td>LE PUBLIC DES REGIONS CONCERNÉES DEVRAIT PRENDRE LES PRECAUTIONS QUI S’IMPOSENT ET SURVEILLER L’EMISSION D’ALERTE SUBSEQUENTES.</td>
<td>LE PUBLIC DES REGIONS CONCERNÉES DEVRAIT PRENDRE LES PRECAUTIONS QUI S’IMPOSENT ET SURVEILLER L’EMISSION D’ALERTE SUBSEQUENTES.</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.000</td>
<td>5</td>
<td>THIS WARNING IS IN EFFECT FROM 2:20 PM TO 4:50 PM EDT.</td>
<td>CETTE ALERTE EST EN VIGUEUR DE 14H20 A 16H50 HAE.</td>
<td>CETTE ALERTE EST EN VIGUEUR DE 14H20 A 16H50 HAE.</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.000</td>
<td>16</td>
<td>SEVERE THUNDERSTORMS HAVE WEAKENED OR HAVE MOVED OUT OF THESE REGIONS.</td>
<td>LES ORAGES VIOLENTS ONT FAIBLI OU ONT QUITTE CES REGIONS.</td>
<td>LES ORAGES VIOLENTS ONT FAIBLI OU ONT QUITTE CES REGIONS.</td>
</tr>
</tbody>
</table>

(anonymous)
Controlled languages - rules

RULE 1:
Write sentences that are shorter than 25 words.

RULE 2:
Write sentences that express only one idea.

RULE 3:
Write the same sentence if you want to express the same content.

RULE 4:
Write sentences that are grammatically complete.

RULE 6:
Write sentences in the active form.

RULE 7:
Write sentences that repeat the noun instead of using a pronoun.
Gisting

Looking for the main idea or most important point in a written or spoken text. Although the translations were served quickly and in all the language pairs that Google Translate offers, some of the formatting and page layout got lost. But if your purpose is gisting, this feature appears to be ideal.

Gisting definition and synonyms | Macmillan Dictionary
www.macmillandictionary.com/dictionary/british/gisting
Gisting

Berliner Morgenpost

In the news: Refugees | In western Berlin | Top News via WhatsApp | All Topics

**UPDATED** REFUGEES

Budapest Station open again - pushing refugees in train

Police have released the train station in Budapest in the morning, the refugees huddled
Gisting for intelligence agencies

3PO. And there are other translation projects in the works, such as the recently announced $5.9 million contract with Raytheon BBN Technologies, to create a real-time English translation of documents, including handwritten notes or images with text on them.

Enter BOLT, which Darpa has asked Congress to fund at $15 million this year. Once developed, BOLT would act something like C-3PO from the Star Wars movies, performing a variety of difficult translation feats for troops in hostile territory.
Gisting for intelligence agencies

- As a first step, select relevant documents from a large collection.
- Interesting documents will then be passed to a human translator.
Integration with speech technologies

Welcome to Skype Translator Preview
Now including two additional spoken languages — German & French — and 50 IM languages

Currently supported for voice/video calls
- Chinese
- English
- French
- German
- Italian
- Spanish

Currently supported for Instant Messages
We currently support 50 languages for Instant Messaging. You can find the full list here.

(http://www.skype.com/en/translator-preview/)
Integration with speech technologies

Broadcast news speech-to-text translation experiments

Sylvain Raybaud    David Langlois    Kamel Smaïli
LORIA - Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy Cedex
givenname.lastname@loria.fr

Development of SRI’s Translation Systems for Broadcast News and Broadcast Conversations

Jing Zheng, Wen Wang, Necip Fazil Ayan
Speech Technology and Research Laboratory, SRI International
{zj,wwang,nfa}@speech.sri.com

A Machine Translation System for Foreign News in Satellite Broadcasting

Teruaki Aizawa, Terumasa Ehara**, Noriyoshi Uratani, Hideki Tanaka,
Naoto Kato, Sumio Nakase*, Norikazu Aruga*, and Takeo Matsuda*
Hand-held devices

- police
- military
- medical
- tourism

(http://www.ectaco.translation.net/)
Hand-held devices

(http://www.ectaco.translation.net/
Tools for translators,
Post-editing

Translation Memory (TM) Match

Matched Sentences  Non-Matched Sentences

Machine Translate  Correct TM Match  Human Translate  Correct MT Output

Edit

Proof

(http://www.languagestudio.com/LanguageStudioDesktop.aspx
Overview

- Introduction
- Applications
- Challenges
- History
- Available resources
- MT paradigms
- MT course
Typology

- Study of cross-linguistic similarities and differences
- Morphology
  - Agglutinative
    - Turkish
  - Fusion
    - Spanish

(http://allthingslinguistic.com/post/50939757945/morphological-typology-illustrations-from)
Typology

- Syntax: e.g., order of verb (V), subject (S) and object (O)

**SVO:**
(German, French, English, Mandarin) → She adores listening to music.

**SOV:**
(Hindi, Japanese) → 彼女は音楽を聴いて大好き。
(she music to listening adores)

**VSO:**
(Irish, Arabic, Biblical Hebrew) → Dúil mhór aici éisteacht le ceol.
(adores she music to listen)
Typology

• Argument structure and linking
  
  - Head-marking:
    - „the man's house“ (English)
  
  - Dependent-marking:
    - „A férfi házában“ „the man house-his“ (Hungarian)
Typology

• Verbs and satellite particles (direction, motion, etc.)

• Verb-framed:
  - Spanish: „La botella salió flotando“ (The bottle exited floating.)

• Satellite-framed
  - English: „the bottle floated out“
Typology

- Pronouns omission
  - Pronoun-drop:
    - English: [I] am reading a book.
    - Spanish: Estoy leyendo un libro.
Typology

- Pronouns omission
  - Referential density
    - Cold: more inferential work to recover antecedents
      - Japanese, Chinese
    - Hot: more explicit and easier
      - Spanish
Lexical

- Homonymy
  - wall (Wand), wall (Mauer)

- Polysemy
  - to know (knowing a fact) : wissen
  - to know (familiarity with a person/location): kennen
Lexical

• Grammar
  - English: „She likes to sing“
  - German: „Sie singt gern."

• Lexical gap
  - „A world view, a philosophy of life“ – Weltanschauung

(http://abbysroad.tumblr.com/post/12947835861/an-incomplete-list-of-english-lexical-gaps)
Other divergences

- Position of adjectives
  - English: „green witch“
  - Spanish: „bruja verde“ - „witch green“
Other divergences

• Cultural aspects, e.g., calendars and dates
  – British English: DD/MM/YY
  – American English: MM/DD/YY
  – Japanese: YYMMDD
Overview

- Introduction
- Applications
- Challenges
- History
- Available resources
- MT paradigms
- MT course
First references to MT

As early as the 17th century by philosophers René Descartes and Gottfried Wilhelm Leibniz

(http://www.biography.com/people/ren-descartes-37613
https://en.wikipedia.org/wiki/Gottfried_Wilhelm_Leibniz)
First references to MT

In 1947, Warren Weaver and Andrew Booth suggested that computers could be used to translate natural languages.
Post WWII:
foreign languages as encrypted English

“One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'”

Georgetown-IBM experiment (1954)

• „[...] human translations were subject to political bias and interference“
• Translation of 60 sentences from Russian into English
• Topic: organic chemistry
• System: six grammar rules and 250 words in the vocabulary

http://www.thelinguafile.com/2013/10/the-georgetown-ibm-experiment-rise-of.html#.Vehb1t93nq5)
Georgetown-IBM experiment (1954)

• Conclusions
  - The problem was solved
  - But semantic disambiguation are impossible to be solved automatically

<table>
<thead>
<tr>
<th>Russian (Romanized)</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mi pyeryedayem misliy posryedstvom ryechyi.</td>
<td>We transmit thoughts by means of speech.</td>
</tr>
<tr>
<td>Vyelyichyina uglia opryedyelyayetsya otnoshenyiyem dlyini dugi k radyiusu.</td>
<td>Magnitude of angle is determined by the relation of length of arc to radius.</td>
</tr>
<tr>
<td>Myezhdunarodnoye ponyimaniye yavlyayetsya vazhnim faktorom v ryeshyenyiyi polyitychyeskix voprosov.</td>
<td>International understanding constitutes an important factor in decision of political questions.</td>
</tr>
</tbody>
</table>
ALPAC report (1966)

- Automatic Language Processing Advisory Committee
- Study of reality of MT
- Conclusions:
  - post-editing not cheaper than full translation
  - Little Russian scientific literature worth to be translated
  - No shortage of human translators
  - No advantage in using machine translation
  - Better fund linguistic research for human translation
- Funding for MT stopped in the US as a consequence

(http://www.hutchinsweb.me.uk/MTNI-14-1996.pdf)
History of MT

• 1970s, first commercial systems
  - Météo
  - Systran
  - Logos
  - METAL
  - Trados
First commercial systems

- 1968: Founded by Dr. Peter Toma
- 1975: Commission of European Communities (CEC)
- 1976: CEC – system from English/France
- 1981: CEC – English/French, French/English, English/Italian
- 1986: Xerox – six target languages
- 1985: SYSTRAN PRO for Windows
- 1997: search engine AltaVista's (today Yahoo's)
- 2006-2007: Google Translate
History of MT

- 1980s, 1990s: interlingual systems
Data-driven methods

- 1980s, Example-based translation

TRAINING:
- Parallel Corpus
  - Sentence Alignment
  - Word Alignment
  - Phrase Extraction
  - Instance Generation

TESTING:
- Input sentences
  - Phrase Extraction
  - Instance Generation
  - Memory-based learning (testing)
  - Decoding
  - Translation
  - Evaluation
  - Score

(http://ilk.uvt.nl/mbmt/pbmbmt/)
Data-driven methods

- Late 1980, Statistical machine translation

"Most state-of-the-art commercial machine translation systems in use today have been developed using a rules-based approach and require a lot of work by linguists to define vocabularies and grammars. Several research systems, including ours, take a different approach: we feed the computer with billions of words of text, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages. We then apply statistical learning techniques to build a translation model."

(http://www.themarysue.com/how-does-google-translate-work/)
Current commercial developers

- Microsoft
- eBay
- IBM
- SDL
- Language Weaver
- Systran
- Google
Overview

- Introduction
- Applications
- History
- Challenges
- Available resources
- MT paradigms
- MT course
Available resources

- Tools
- Parallel corpora
Tools

- Natural language processing (NLP) tools:
  - Tokenization, parsing, named-entity recognition

- MT tools:
  - GIZA++: IBM's word-based models
  - Moses, Thot: phrase-based models

- MT evaluation tools:
  - BLEU, METEOR
Parallel corpora

- LDC, Gigaword

[link to LDC catalog](https://catalog.ldc.upenn.edu/LDC2011T11)
[link to LDC catalog](https://catalog.ldc.upenn.edu/LDC2011T07)
Parallel corpora

- Europarl

  - source release (text files), 1.5 GB
  - tools (preprocessing tools and sentence aligner only), 8.6 KB
  - parallel corpus Danish-English, 179 MB, 04/1996-11/2011
  - parallel corpus German-English, 189 MB, 04/1996-11/2011

(http://www.statmt.org/europarl/)
## Parallel corpora

- **Acquis Communautaire**

<table>
<thead>
<tr>
<th>Language</th>
<th>ISO Code</th>
<th>N° of Texts</th>
<th>Text body</th>
<th>Signature</th>
<th>Annex</th>
<th>Total N° Words (Text + Signature + Annex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>7983</td>
<td>5979261</td>
<td>38479314</td>
<td>749</td>
<td>609441</td>
<td>2100301</td>
</tr>
<tr>
<td>da</td>
<td>7939</td>
<td>6548461</td>
<td>44444011</td>
<td>825</td>
<td>691894</td>
<td>1599456</td>
</tr>
<tr>
<td>de</td>
<td>7914</td>
<td>6576633</td>
<td>47047334</td>
<td>831</td>
<td>571928</td>
<td>1506847</td>
</tr>
<tr>
<td>el</td>
<td>7782</td>
<td>7377316</td>
<td>47715936</td>
<td>948</td>
<td>559487</td>
<td>1628451</td>
</tr>
<tr>
<td>en</td>
<td>7972</td>
<td>7512013</td>
<td>45150120</td>
<td>942</td>
<td>667978</td>
<td>1752545</td>
</tr>
<tr>
<td>es</td>
<td>7809</td>
<td>7964255</td>
<td>48281455</td>
<td>1020</td>
<td>709279</td>
<td>1832745</td>
</tr>
<tr>
<td>et</td>
<td>7944</td>
<td>4925361</td>
<td>38603952</td>
<td>620</td>
<td>439184</td>
<td>1819226</td>
</tr>
<tr>
<td>fi</td>
<td>7735</td>
<td>5134294</td>
<td>43705813</td>
<td>664</td>
<td>565226</td>
<td>1180877</td>
</tr>
<tr>
<td>fr</td>
<td>7862</td>
<td>7812577</td>
<td>45609935</td>
<td>994</td>
<td>673061</td>
<td>1726720</td>
</tr>
<tr>
<td>hu</td>
<td>7489</td>
<td>5391810</td>
<td>40601868</td>
<td>720</td>
<td>539967</td>
<td>188746</td>
</tr>
<tr>
<td>it</td>
<td>7872</td>
<td>7264126</td>
<td>46792286</td>
<td>923</td>
<td>707467</td>
<td>1704221</td>
</tr>
<tr>
<td>lt</td>
<td>7966</td>
<td>5386359</td>
<td>39936370</td>
<td>676</td>
<td>625365</td>
<td>1948354</td>
</tr>
<tr>
<td>lv</td>
<td>7980</td>
<td>5656335</td>
<td>39290110</td>
<td>709</td>
<td>461736</td>
<td>2011426</td>
</tr>
<tr>
<td>mt</td>
<td>7639</td>
<td>7203858</td>
<td>43919881</td>
<td>947</td>
<td>505324</td>
<td>2288013</td>
</tr>
<tr>
<td>nl</td>
<td>7882</td>
<td>7394645</td>
<td>47699508</td>
<td>931</td>
<td>712255</td>
<td>1710041</td>
</tr>
<tr>
<td>pl</td>
<td>7968</td>
<td>5974605</td>
<td>43160945</td>
<td>750</td>
<td>668248</td>
<td>2070687</td>
</tr>
<tr>
<td>pt</td>
<td>7848</td>
<td>7851904</td>
<td>47225710</td>
<td>1001</td>
<td>648180</td>
<td>1838833</td>
</tr>
<tr>
<td>ro</td>
<td>5792</td>
<td>5122354</td>
<td>33681450</td>
<td>884</td>
<td>402929</td>
<td>4047393</td>
</tr>
<tr>
<td>sk</td>
<td>5278</td>
<td>3911895</td>
<td>26077956</td>
<td>741</td>
<td>413511</td>
<td>1381471</td>
</tr>
<tr>
<td>sl</td>
<td>7984</td>
<td>5989322</td>
<td>37844883</td>
<td>750</td>
<td>573052</td>
<td>2153138</td>
</tr>
<tr>
<td>sv</td>
<td>7731</td>
<td>6472717</td>
<td>42990411</td>
<td>837</td>
<td>560188</td>
<td>1424887</td>
</tr>
<tr>
<td>Average</td>
<td>7,030</td>
<td>6,353,410</td>
<td>42,340,925</td>
<td>831</td>
<td>585,947</td>
<td>1,880,338</td>
</tr>
</tbody>
</table>

([Link to the original document](http://optima.jrc.it/Acquis/JRC-Acquis.2.2/doc/README_Acquis-Communautaire-corpus_JRC.html))
Resumo

VIEIRA, Marineide R. e CHIAVEGATO, Luiz G. 

No estudo da biologia de Polyphagotarsonemus latus em limão Siciliano, foram utilizados potes plásticos circulares com capacidade de 250 ml, contendo areia esterilizada como suporte para dois frutos novos com aproximadamente 2,0 cm de diâmetro. O ensaio foi conduzido a 27,1 ± 0,5°C, umidade relativa de 67,6 ± 1,3% e fotofase contínua. O período de ovo a adulto durou 3,7 ± 0,1 dias para fêmeas e 3,6 ± 0,1 dias para machos, com sobrevivência de 100%. Após um período de pré-oviposição de 1,0 ± 0,2 dias, as fêmeas depositaram 5,6 ± 0,5 ovos por dia durante 10,5 ± 0,9 dias, totalizando 58,9 ± 6,7 ovos por fêmea. A longevidade foi de 13,4 ± 1,0 dias para fêmeas e 12,0 ± 2,4 dias para machos. A razão intrínseca de aumento (rm) foi de 0,359, a razão finita de aumento (l) de 1,43 indivíduos por fêmea por dia, o tempo médio de uma geração (T) de 10,34 dias e a taxa líquida de reprodução (Ro) de 41,0.

Palavras-chave: Ácaro branco; desenvolvimento biológico; tabela de vida de fertilidade; taxa líquida de reprodução.

Resumo

VIEIRA, Marineide R. e CHIAVEGATO, Luiz G. 

In the study of the biology of Polyphagotarsonemus latus (Banks) on lemon var. Siciliano (Citrus limon Burm) round plastic pots (250 ml) containing sterilized sand were used as support for two 2cm-diameter new fruits. The assay was carried out at 27,5 ± 0,5°C, relative humidity of 67,6 ± 1,3% and constant photophase. The duration of immature phases was 3,7 ± 0,1 days for females and 3,6 ± 0,1 days for males, with 100% survival. After a pre-oviposition period of 1,0 ± 0,2 days, the females deposited 5,6 ± 0,5 eggs per day during 10,5 ± 0,9 days, i.e., 58,9 ± 6,7 eggs per female. The longevity was 13,4 ± 1,0 days for females and 12,0 ± 2,4 days for males. The intrinsic rate of increase (rm) was 0,359, finite rate of increase (l) 1,43 individual per female per day, mean generation time (T) 10,34 days and net reproductive rate (Ro) 41,0.

Palavras-chave: Broad mite; biological development; life table of fertility; net reproductive rate.
Evaluation campaigns

NIST Open Machine Translation 2015 Evaluation (OpenMT15)

Highlights

- Evaluation on informal data genres (SMS/Chat, Conversational Telephone Speech) for Arabic-to-English and Chinese-to-English
- Inclusion of audio input track
- Explore common MT measurement techniques on these informal data genres
Evaluation campaigns

IWSLT 2015, International Workshop on Spoken Language Translation

- Automatic speech recognition (ASR), i.e. the conversion of a speech signal into a transcript,
- Machine translation (MT), i.e. the translation of a polished transcript into another language,
- Spoken language translation (SLT), i.e. the conversion and translation of a speech signal into a transcript in another language.
Evaluation campaigns

ACL 2016
FIRST CONFERENCE ON
MACHINE TRANSLATION (WMT16)

- Czech-English
- German-English
- Romanian-English
- Finnish-English
- Russian-English
- Turkish-English

Shared Task: Biomedical Translation Task

- English-French and French-English
- English-Spanish and Spanish-English
- English-Portuguese and Portuguese-English

(http://www.statmt.org/wmt16/)
Overview

- Introduction
- Applications
- Challenges
- History
- Available resources
- MT paradigms
- MT course
MT paradigms

Machine Translation

Knowledge-based MT
- Rule-based MT

Data-driven MT
- Example-based MT
- Statistical MT
- Neural MT
Rule-based MT

Rule-based MT

• Apertium (https://www.apertium.org)
Example-based MT

入力:
ウイスキーはオオムギから製造される

翻訳用例データベース

出力:
whisky is produced from barley

(http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?plugin=attach&refer=KUROHASI-KAWAHARA-LAB&openfile=EBMT.png)
Example-based MT

- Cunei (http://cunei.sourceforge.net/)
- KyotoEBMT (http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KyotoEBMT)
Statistical MT

[Diagram showing the process of statistical machine translation with nodes for translation model, language model, parallel corpus, and monolingual corpus leading to English output.]

Statistical MT

- Moses (http://www.statmt.org/moses/)
- Cunei (http://cunei.sourceforge.net/)
Neural MT

Figure 2: The very first neural machine translation system.

Neural MT

- LISA (http://104.131.78.120/)
Overview

- Introduction
- Applications
- Challenges
- History
- Available resources
- MT course
MT course – what to expect from me

- Overview on MT methods
- Supervision of the projects
- Be available by email and in the office (Villa room 0.01)
MT course – what I expect from you

- Presence and participation in the lecture (not controlled)
- Take part in a project (team or individual)
- Take part in the final exam
Project

- Teams of 2/3 students
- „Take part“ in one of the translation challenges at WMT‘16 (http://www.statmt.org/wmt16/)
  - News
  - IT-domain
  - Biomedical
- Presentation of preliminary and final results
- Submission of a 3-pages report
- Source code in GitHub or similar
Project

• Flexible...
  • „Any“ translation task (first-come, first-served)
  • Any language pair
  • Any MT paradigm
  • Any NLP/MT tools
Project

- ...but with some requirements
  - Integration of domain-specific resources
  - Training on out-of-domain corpora (talk to other teams)
  - Evaluation of official test datasets (last year’s test data)
Project

- Mail to me (mariana.neves@hpi.de):
  - Team members
  - WMT translation task(s)
  - Language pair(s)
  - Host of the project (GitHub, etc)
## Lectures

(Program is subject to change)

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Oct 17, 2016</td>
<td>Introduction to Machine Translation</td>
</tr>
<tr>
<td>2</td>
<td>Oct 24, 2016</td>
<td>Words, sentences and corpora</td>
</tr>
<tr>
<td>3</td>
<td>Oct 31, 2016</td>
<td>(Reformationstag)</td>
</tr>
<tr>
<td>4</td>
<td>Nov 7, 2016</td>
<td>Word alignment</td>
</tr>
<tr>
<td>5</td>
<td>Nov 14, 2016</td>
<td>Statistical word-based models</td>
</tr>
<tr>
<td>6</td>
<td>Nov 21, 2016</td>
<td>Statistical phrase-based models</td>
</tr>
<tr>
<td>7</td>
<td>Nov 28, 2016</td>
<td>Language model</td>
</tr>
<tr>
<td>8</td>
<td>Dec 5, 2016</td>
<td>Neural MT</td>
</tr>
<tr>
<td>9</td>
<td>Dec 12, 2016</td>
<td>(no lecture? - to be confirmed)</td>
</tr>
<tr>
<td>10</td>
<td>Jan 2, 2017</td>
<td>Mid-term presentation of projects</td>
</tr>
<tr>
<td>11</td>
<td>Jan 9, 2017</td>
<td>Decoding</td>
</tr>
<tr>
<td>12</td>
<td>Jan 16, 2017</td>
<td>Evaluation</td>
</tr>
<tr>
<td>13</td>
<td>Jan 23, 2017</td>
<td>Rule-based MT</td>
</tr>
<tr>
<td>14</td>
<td>Jan 30, 2017</td>
<td>Memory-based MT</td>
</tr>
<tr>
<td>15</td>
<td>Feb 6, 2017</td>
<td>Final presentation of projects</td>
</tr>
<tr>
<td>16</td>
<td>Feb 13, 2017</td>
<td>Final exam</td>
</tr>
</tbody>
</table>
Grading

- 60% Project
  - Commitment, implementation, presentation, report
  - Each team member should present in either of the two appointments (mid-term or final)

- 40% Final exam
Course books

- Statistical Machine Translation
  - Philipp Koehn

- Machine Translation
  - Pushpak Bhattacharyya
Course books

- Speech and Language Processing (Chapter 25)
  - Daniel Jurafsky and James H. Martin
Course books (advanced topics)

- Learning Machine Translation
  - Edited by Cyril Goutte, Nicola Cancetta, Marc Dymetman
# Workshop Papers

<table>
<thead>
<tr>
<th>pdf</th>
<th>bib</th>
<th>Title</th>
<th>Authors</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdf</td>
<td>bib</td>
<td>Cross-language Projection of Dependency Trees with Constrained Partial Parsing for Tree-to-Tree Machine Translation</td>
<td>Yu Shen, Chenhui Chu, Fabien Cromieres and Sadao Kurohashi</td>
<td>pp. 1–11</td>
</tr>
<tr>
<td>pdf</td>
<td>bib</td>
<td>Improving Pronoun Translation by Modeling Coreference Uncertainty</td>
<td>Ngoc Quang Luong and Andrei Popescu-Belis</td>
<td>pp. 12–20</td>
</tr>
<tr>
<td>pdf</td>
<td>bib</td>
<td>Modeling verbal inflection for English to German SMT</td>
<td>Anita Ramm and Alexander Fraser</td>
<td>pp. 21–31</td>
</tr>
<tr>
<td>pdf</td>
<td>bib</td>
<td>Modeling Selectional Preferences of Verbs and Nouns in String-to-Tree Machine Translation</td>
<td>Maria Nadejde, Alexandra Birch and Philipp Koehn</td>
<td>pp. 32–42</td>
</tr>
<tr>
<td>pdf</td>
<td>bib</td>
<td>Modeling Complement Types in Phrase-Based SMT</td>
<td>Marion Weller-Di Marco, Alexander Fraser and Sabine Schulte im Walde</td>
<td>pp. 43–53</td>
</tr>
<tr>
<td>pdf</td>
<td>bib</td>
<td>Alignment-Based Neural Machine Translation</td>
<td>Tamer Alkhouli, Gabriel Bretschner, Jan-Thorsten Peter, Mohammed Hethnawi, Andreas Guta and Hermann Ney</td>
<td>pp. 54–65</td>
</tr>
<tr>
<td>pdf</td>
<td>bib</td>
<td>Neural Network-based Word Alignment through Score Aggregation</td>
<td>Joël Legrand, Michael Auli and Ronan Collobert</td>
<td>pp. 66–73</td>
</tr>
<tr>
<td>pdf</td>
<td>bib</td>
<td>Using Factored Word Representation in Neural Network Language Models</td>
<td>Jan Niehues, Thanh-Le Ha, Eunah Cho and Alex Waibel</td>
<td>pp. 74–82</td>
</tr>
</tbody>
</table>